DOCUMENT RESUME

ED 390 917 TM 024 320

AUTHOR Goldberg, Gail Lynn; Michaels, Hillary
TITLE Same-Scorer Judgments on Multiple Content Area Items in Integrated Performance Assessment.
PUB DATE 22 Apr 95
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; *Evaluators; Grade 3; Grade 8; *Interrater Reliability; Language Proficiency; Mathematics Achievement; *Performance Based Assessment; Reading Achievement; Scores; *Scoring; *Test Content; Test Results; *Training
IDENTIFIERS *Maryland School Performance Assessment Program

ABSTRACT

Preliminary data was gathered to guide subsequent research that will shape training procedures and scoring practice for performance assessment activities that integrate multiple content areas. Content area integration is a key feature of many of the tasks in the Maryland School Performance Assessment Program (MSPAP), a large-scale assessment of all students at grades 3, 5, and 8 in reading, language use, mathematics, science, and social studies. Content area assessment in the MSPAP has required simultaneous scoring, the use of a single scoring tool to make one judgment on two or more outcomes, and sequential scoring, the use of different tools to make consecutive judgments. At the end of the 1993 MSPAP, 10 scorers each scored 100 responses from grade 3 or grade 8. Each scorer used three different scoring tools, a writing rule, a language-in-use rule, and a social studies activity key on the constructed response items. Results supported the use of a single scorer to score different content areas, and highlighted areas for further research on rater training and scoring multiple content areas. Five appendixes describes the scoring rules. (Contains three tables.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 390 917

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GAIL LYNN GOLDBERG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

# Same-Scorer Judgments on Multiple Content Area Items in Integrated Performance Assessment

Gail Lynn Goldberg and Hillary Michaels

Maryland State Department of Education

Paper presented at the annual meeting of the American Educational Research Association

San Francisco, California

22 April 1995

TM024320



BEST COPY AVAILABLE

# Same-Scorer Judgments on Multiple Content Area Items in Integrated Performance Assessment

Gail Lynn Goldberg and Hillary Michaels
Maryland State Department of Education

## Objectives

Performance assessment almost invariably requires judgment-based scoring; generally guided by specific criteria, readers make score decisions about the level of performance demonstrated in student responses. Typically, when multiple content areas (e.g., writing, reading, mathematics) are being assessed, different readers make judgments about performance in each of those areas. This practice may be neither appropriate nor necessary, however, when multiple measures are being obtained from the same activity or task. Multiple judgments have long been made by the single reader who applies an analytical checklist to assign sometimes divergent scores for different traits, as well as by the teacher who gives the familiar "split grade" (e.g., A/B- for content/mechanics). Particularly given recent trends in content area integration in both instruction and assessment, the ability to obtain valid, reliable data based on same-scorer judgments must be questioned. The purpose of this study was to gather preliminary data to guide subsequent research which will shape training procedures and scoring practice for

performance assessment activities that integrate multiple content areas (CAs).

## Perspectives

Recently evolving standards for teaching and learning require "interdisciplinary curricula that engage students in integrative ways of thinking and learning" (Education Week, 1995). Concurrently, the last few years have witnessed an increase not only in the use of performance assessments in a variety of content areas, but in the integration of those content areas in performance tasks that mirror more authentic and complex tasks which students will face in adult life and in the world of work (Birukoff, Ferrara, Householder and Goldberg, 1994). This sort of content area integration is a key feature of many of the tasks that comprise the Maryland School Performance Assessment Program (MSPAP), a large-scale assessment of all students at Grades 3, 5, and 8, in reading, writing/language in use, mathematics, science, and social studies. MSPAP tasks are comprised of related activities (or items), each receiving one or more scores on one or more Maryland Learning Outcomes. In MSPAP, content area integration has required two different scoring strategies: **simultaneous scoring**, or the use of a single scoring tool (score scale and descriptive criteria) to make one judgment encompassing two or more outcomes, and **sequential scoring**, or the use of different scoring tools to make consecutive and sometimes different judgments about performance on different outcomes. Cognizant of the possibility that judgments in one area may

affect those in others, we have until now generally assigned application of different scoring tools for the same activity (or item) to different readers, with the exception of sequential scoring for writing (W) and language in use (LU). As content area integration has become a greater feature of instruction and assessment in our state, the need for research to confirm or refute this practice became apparent.

## Methods

At the end of operational scoring of the 1993 assessment, five readers were selected for each of two grade levels (3 and 8) from among the larger pool of trained readers (all of whom are Maryland educators) who had worked on that project. As in the case of operational scoring team assignments, those scoring the grade 3 study sample were generalists (typically, elementary school teachers) and those scoring grade 8 were specialists (e.g., either English language arts teachers or social studies teachers). However, in no instances had any of the readers selected for the study scored the same items during operational scoring (and thus none had already had concentrated training on scoring for a particular content area). All ten readers had average records for scoring in these CAs in the range of 80-85% exact agreement with pre-established "true" scores (86% being the project-wide exact agreement rate in 1993), based on twice-weekly validity packets administered during operational scoring. At each grade level, these readers were trained on three different scoring tools: a writing rule, a language in use rule, and a

social studies activity-specific key (See Appendices 1-5). "Rules" are simply brief rubrics, or generic score scales accompanied by score point descriptors, while the "activity-specific keys" are score scales accompanied by descriptors that are unique to the given activity to be scored. In each case, the study item was designed to elicit a constructed response (poem or paragraph) to be scored for each of the three content areas. These items are designated LWP (limited writing process) items to distinguish them from prompts for extended, essay-length responses (EWP, or extended writing process items) which are subject to peer response and revision during the assessment. These LWP items were selected from among the pool of items requiring sequential scoring because of their common social studies focus, and because all scoring tools used the same score scale (0-2). CA-specific training materials were used (the same materials as those which had been used for operational scoring training) but instances were highlighted during scoring training when CA scores (W, LU, and SS) should be discrepant for a particular training sample. After training, all readers for each grade level scored all 100 study responses; these responses, selected to represent all score points, had been organized for scoring into five randomly assigned packets of 20. Although the order of decisions on the monitor sheets was SS-W-LU, readers were not required to assign scores for each student response in this order.

We calculated percent agreement with the "standard" (given

operational score) and performed an analysis of variance to determine any differences between raters. In operational scoring, readers must maintain a minimum exact agreement rate of 70% (against "true" scores pre-established by a team of highly experienced readers). Therefore, this same benchmark was used initially as one means of determining whether we had approached, met, or exceeded a satisfactory agreement rate to permit a single reader to make multiple CA judgments, instead of utilizing multiple readers.

In addition, score data were analyzed to determine if the same reader could make judgments which maintained the same relationships (discrepancy or consistency) among all three areas as the ones which were identified between scores assigned by different readers. Because the data are ordinal rather than on a continuous scale, we ran an analysis of variance using ranked scores (NPAR1WAY) in SAS. Data were analyzed in "batches." Each batch corresponded to one of the five randomly assigned sets of student responses scored by each reader. Batches were considered independent of each other to minimize any possible order effect.

Data Source

One hundred responses for each grade level (3rd and 8th) were purposefully selected from the larger pool so that all blank responses were removed and all score points were represented (some consistent across CAs and others discrepant). All responses had been previously scored by other readers trained either in W and LU, or on the SS tool. These scores were entered

as the standard for purposes of comparison, although they did not represent consensus judgment on each response of a larger group of readers, as "true" scores often do. Operational scores were assigned randomly by one reader on a team of 16-18 members, on average. Thus, the mean score for the standard on each packet, which was used for some of our analyses, represented multiple reader judgments as did the study mean.

## Results and Discussion

Overall, in both Grades 3 and 8, the exact and adjacent agreement rates taken together was in the high 90s in all areas. Across all CAs, the average exact agreement rate was 70% in Grade 3 but only 60% in Grade 8. This lower percentage in Grade 8 may be attributed to readers using the whole score range (0-2) while in Grade 3, we observed more 0-1 decisions. Since the standard was based on a single reader's judgment, however, and was not representative of a "true score, we recognized that strict comparison with conventional agreement "targets" was not appropriate or adequate by itself to confirm or reject the feasibility of same-scorer judgments.

The quality of judgment on performance assessment is subject to a variety of reader effects such as rater severity, halo effect, central tendency and restriction of range (Engelhard, Jr., 1994). The effects which are likely to occur when a given reader scores multiple responses to the same item can also be observed when readers score the same item with multiple sets of criteria (See Tables 1 and 2). Evidence of these effects in the

same-scorer study appear at grade 3 but not at grade 8. At grade 3 (See Table 1), readers tended to score with the same degree of severity (e.g., scoring high or scoring low) within each CA, and between W and SS. This would suggest that there is a "blending" of judgments in these two areas, while readers were able to apply LU criteria independent of these areas. Further, in 3rd grade, the tendency to vary from the standard (with readers #1-5 always scoring more harshly, on average) is consistent across CAs. The relationship among CAs in terms of the average percent of the maximum score (2) over a batch changed such that while SS was always the most difficult CA, LU rather than W was the least difficult when the same reader scored all areas. In 8th grade, however, the tendency to vary from the standard is not evident (See Table 2). The relationship among the CAs remains the same as well.

We defined "approaching the standard" as being within .10 of the mean original scores (the mean standard scores) for each batch. Results indicate that the standard was never approached in Grade 3 but was approached for all areas in Grade 8 (See Table 3). The least agreement with the standard occurred in Grade 3 W (.22) closely followed by SS (.20). The greatest agreement with the standard was .01 in Grade 8 Writing, followed by SS (.04). The congruency between the standard and mean study readers' scores in 8th grade is all the more powerful because of evidence

of severity in judgment by one reader (ID #4)[1] consistently. Across both batches and CAs, reader #4 is always lower than her counterparts. Had that reader's scores been disregarded, along with the W scores of another aberrant reader (#2--the most lenient reader), the congruency would be nearly perfect, in fact, for SS and W. However, LU would still retain a difference of .10 because more of these readers were lenient in relation to the standard. Looking at both Grade 3 and 8 score data, it appears that most readers score more leniently for LU than for the other two areas when they are scoring all CAs.

The ability to approach the standard is especially important in light of the fact that MSPAP was designed to provide school and system level data, not individual student scores. Thus, while exact inter-rater agreement rates were low for Grade 8, the fact that batch mean scores approached the standard suggests that it is indeed defensible to use same-scorer judgments at least for this particular assessment program.

Other evidence of reader consistency came from the ANOVA results. Only three (out of thirty) of the ANOVAs were significant at the .05 level. These were Grade 3 batch 3 LU (reader #4), Grade 3 batch 3 W (again, reader #4), and Grade 8 batch 1 W (readers #4 and #5). This implies considerable consistency among readers and suggests that the training protocol for multiple CAs was successful.

---

[1] Rater number in Tables 1 and 2 refers to reader ID number, not to order of raters' readings of "batched" responses.

Paradoxically, the lesser ability of Grade 3 readers to make independent judgments may be related to their instructional status as generalists, compared to Grade 8 readers, who are CA specialists. That is, these readers function most effectively when trained on CA-specific criteria and held to a focus on that one area. Otherwise, their tendency to "generalize" kicks in. In contrast, the Grade 8 readers are trained and experienced in operating with specific criteria applied independent of others. They are able to look through one "lens," and recognize that others are necessary in order to make valid judgments in different areas.

Educational Importance and Practical Implications

We frequently hear the adages that "good assessment models good instruction" and that "if you test it, it will be taught" (cf. Wiggins, 1993, p. 5 for similar aphorisms). There are strong instructional arguments, therefore, for designing assessments to mirror the content area integration that is increasingly becoming a hallmark of good instruction. In so doing, however, there are clearly a host of scoring-related issues to be addressed, foremost among them the necessity or advisability of independent decision-making on performance in multiple content areas based on the same response(s). From an instructional perspective, teachers' increased understanding of, and ability to, identify discrepant degrees of proficiency will help teaching and learning based on multiple content areas. From a practical perspective, we may be able to make more informed

decisions that impact project time and cost of scoring.

In order to make such decisions, it seems advisable for the time being to take a conservative stance. Therefore, whenever possible, MSPAP is still maintaining separate scoring teams to assign scores for responses to the same item which yield multiple CA measures although our preliminary data on same scorer judgments are promising. Given cost constraints and the complex logistics of booklet and score sheet flow, however, it is proving far easier to manipulate team designs to separate W/LU from other CA score decisions than to assign those others (e.g., SS and science, or science and mathematics) to different scoring teams. With MSPAP design cautiously moving in the direction of more frequent, and more complex, integration, the need for continued study is unquestionable.

In the next phase of this study, we intend to explore different training protocols to further improve agreement rates when scoring multiple CAs and to obtain more information on the readers involved, both through background questionnaires and think-aloud protocols during scoring. We anticipate focusing on activities which combine different sets of CA outcomes and purposes for writing. Given the increased number of LWP items in the 1995 edition, we have more options and have selected the following items for the study sample:

Grade 3: science + writing to inform

social studies + writing to persuade

Grade 5: mathematics + writing to inform

science + writing to persuade

social studies + writing to inform

Grade 8: mathematics + writing to persuade

science + writing to inform

social studies + writing to persuade

In addition, we will include several items not scored for writing but for two CAs (e.g., reading and mathematics, reading and science, science and mathematics). Under current practice, these are sometimes scored by the same reader; therefore, using reverse procedures, we will train two separate teams and compare separate and same-scorer judgments on responses to these items.

Particularly because of instructional concerns about the confounding of reading (R) measures through written response, we would like to also investigate the relationship between R and W scores on the same item(s). Unfortunately, various psychometric exigencies make this more difficult at present. Reading tasks are usually designed to be coupled with a measure for writing derived from an extended, prompted response rather than a LWP item. Since student absence on any day when a given CA is measured requires that record to be dropped when determining scale scores, in 1995 and beyond extended writing is unlikely to be scored for any other CAs. We recognize, however, that this is an issue that remains to be addressed.

Indeed, there are a considerable number of research questions related to scoring integrated assessment that beg attention. Some of these include:

- How does the order of score decisions on different CAs effect accuracy?
- What impact does the particular purpose for writing (e.g., to inform, persuade, or express personal ideas) have on scoring for multiple CAs? Is it easier or more difficult to keep decisions distinct when scoring writing for different purposes?
- How does reader background affect score accuracy when making decisions in multiple CAs? That is, do readers perform more effectively when they are generalists than when they are specialists in a given content area?
- What effect on accuracy, if any, is there when the score scales used are not of the same range? For example, are readers more or less accurate when making several decisions using the same scale (e.g., 0-2) than when using various scales (e.g., 0-2 and 0-3)?
- Are some CAs more "discourse-friendly" than others? That is, is multiple CA scoring better suited to those content areas in which writing is a more customary response mode like social studies, and less well suited to an area like mathematics, in which extended written response has only recently become a part of instruction?

Many other questions as well may derive from our initial investigation, and hopefully from the dialogue based on it that will ensue. We urge that others interested in state-of-the-art performance assessment join in investigating the scoring

implications of content area integration to ensure valid, reliable assessment that supports, and is in concert with, exemplary instruction.

## References


Birukoff, K., Ferrara, S., Householder, D., and Goldberg, G. (1994, June). Content area integration: implications for large-scale performance assessment and instruction. Paper presented at the CCSSO National Conference on Large-Scale Assessment. Albuquerque, NM.

Engelhard, G. Jr. (1994). Examining rater errors. Journal of Educational Measurement, 31(2), 93-112. Engelhard, G. Jr. (1994). Examining rater errors. Journal of Educational Measurement, 31(2), 93-112.

Goldberg, G. (1993). Scoring MSPAP: A Teachers Guide. Baltimore, MD: MD State Department of Education.

"NSDC and NAESP Publish Elementary School Staff Development Standards," Education Week, 14 (25), 7.

Wiggins, Grant P. (1993). Assessing student performance: exploring the purpose and limits of testing. San Francisco, CA: Jossey-Bass.

Appendix 1. Writing to Express Personal Ideas Rule

SCORING RULE: WRITING TO EXPRESS PERSONAL IDEAS

2 = Consistently addresses audience's needs by presenting personal ideas in a complete, well-developed whole. Text is uniformly organized, and language choices often enhance the text and are appropriate to the literary form.

1 = Sometimes addresses the audience's needs in an incomplete or partially developed whole. Text is generally organized, and language choices sometimes enhance the text and may sometimes be appropriate to the literary form.

0 = Rarely or never addresses audience's needs by failing to present personal ideas in a complete, well-developed whole. Text is often disorganized, and language choices seldom, if ever, enhance the text and are often inappropriate to the literary form.

Appendix 2. Writing to Persuade Rule

SCORING RULE: WRITING TO PERSUADE

2 = Consistently addresses audience's needs by identifying a clear position and fully supporting or refuting that position with relevant information. Text is uniformly organized, and language choices often enhance the text.

1 = Sometimes addresses audience's needs by identifying a somewhat clear position and partially supporting or refuting that position with relevant information. Text is generally organized, and language choices sometimes enhance the text.

0 = Rarely or never addresses the audience's needs by failing to identify a clear position or failing to adequately support or refute a position that has been identified. Text lacks organization, and language choices seldom, if ever, enhance the text.

Appendix 3. Language in Use Rule

SCORING RULE: LANGUAGE IN USE

2 = Consistently uses word and sentence order and language choices to express meaning with style and tone. Text conveys uniform impression of correctness* and any errors that are present represent risk-taking.

1 = Sometimes uses word and sentence order and language choices to express meaning with style and tone. Text generally conveys impression of correctness* and errors may or may not represent risk-taking.

0 = Rarely or never uses word and sentence order and language choices to express meaning with style and tone. Text appears error-ridden.

*correct usage, punctuation, spelling, and capitalization

Appendix 4. Activity description and activity-specific scoring key (social studies) for 3rd grade activity

Description of activity:

Students are given an untitled poem about [a location]* and asked to add information to the poem to help others understand where that location is. They are asked to name the location, describe it, and include either a landform or body of water associated with it.

Activity-specific scoring key:

The response gives evidence of an understanding of geographic concepts.

2 = The response names and provides at least one description of the [location] plus a landform or body of water that is related to that [location]

1 = The response includes only a partial (partially complete or correct) description

0 = Other

Answer Cue:

| NATURAL LANDFORM | NATURAL BODIES OF WATER |
|---|---|
| mountain (mountain chain) | lake |
| plain | river |
| volcano | ocean |
| island | stream |
| peninsula | bay |
| plateau | gulf |
| valley | spring |
| hills | |

Note: References to landform or body of water must be accurate

*Specific details of this activity have been removed to maintain task security

Appendix 5. Activity description and activity-specific scoring key (social studies) for scoring 8th grade activity

Description of activity:

Students are asked to consider all the sources they read as part of the task, and then decide whether or not the U.S. should place limits on the use of [a particular natural resource]. They are cued to think about the short-term and long-term economic effects of their decision. They are then asked to write a letter to the editor of the local newspaper stating and supporting their position, and including at least one short- and one long-term economic effect.

Activity-specific key:

The response gives evidence of an understanding of the historical development and current status of economic principles, institutions and processes needed to be effective citizens, consumers, and workers in American society.

2 = The response gives thorough evidence by stating a position and describing at least one correct short-term AND one correct long-term economic consequence of the choice.

1 = The response gives adequate evidence by stating a position and describing at least one correct short-term OR one correct long-term economic consequence of the choice.

0 = Other

Answer Cue:

Possible short-term economic consequences (if position FOR limits is presented):

- rise in prices of [particular natural resource]
- fall in demand for [particular natural resource]
- fall in profits from production of [particular natural resource]
- fall in production of product made from [particular natural resource]
- any appropriate consequence, including very specific ones

Possible long-term economic consequences (if position FOR limits is presented):

- development of new technology

- use of (more costly) alternative products
- smaller, more efficient [homes, stores, etc.]
- use of alternative resources
- change or reduction in size/scope of businesses involved in producing or distributing products made from [particular natural resource]
- any appropriate consequence, including very specific ones

Or any other feasible responses, based on position

*Specific details of this activity have been removed to maintain task security

G31:aera95dr

Table 1

*Grade 3 rater averages by batch for social studies, writing, and language usage.*

Social Studies

| Scorer ID | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Average |
|---|---|---|---|---|---|---|
| Standard | 0.35 | 0.50 | 0.60 | 0.60 | 0.50 | 0.51 |
| Rater 1 | 0.21 | 0.20 | 0.50 | 0.30 | 0.30 | 0.30 |
| Rater 2 | 0.36 | 0.25 | 0.25 | 0.30 | 0.25 | 0.28 |
| Rater 3 | 0.26 | 0.40 | 0.60 | 0.45 | 0.50 | 0.44 |
| Rater 4 | 0.26 | 0.35 | 0.50 | 0.25 | 0.40 | 0.35 |
| Rater 5 | 0.10 | 0.10 | 0.45 | 0.20 | 0.25 | 0.22 |

Writing

| Scorer ID | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Average |
|---|---|---|---|---|---|---|
| Standard | 0.63 | 0.45 | 1.00 | 0.80 | 0.75 | 0.73 |
| Rater 1 | 0.42 | 0.35 | 0.70 | 0.35 | 0.55 | 0.47 |
| Rater 2 | 0.63 | 0.50 | 0.60 | 0.75 | 0.65 | 0.63 |
| Rater 3 | 0.63 | 0.50 | 1.05 | 0.70 | 0.85 | 0.75 |
| Rater 4 | 0.21 | 0.30 | 0.25 | 0.25 | 0.55 | 0.31 |
| Rater 5 | 0.31 | 0.20 | 0.55 | 0.60 | 0.45 | 0.42 |

Language Usage

| Scorer ID | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Average |
|---|---|---|---|---|---|---|
| Standard | 0.58 | 0.35 | 1.05 | 0.75 | 0.65 | 0.68 |
| Rater 1 | 0.52 | 0.65 | 0.90 | 0.45 | 0.55 | 0.61 |
| Rater 2 | 0.52 | 0.55 | 0.55 | 0.55 | 0.60 | 0.55 |
| Rater 3 | 0.47 | 0.50 | 0.80 | 0.60 | 0.85 | 0.64 |
| Rater 4 | 0.36 | 0.25 | 0.40 | 0.25 | 0.45 | 0.34 |
| Rater 5 | 0.63 | 0.65 | 0.80 | 0.60 | 0.85 | 0.71 |

Note: Batches represent rater averages (n = 20) on scales of 0 - 2.





PRELIMINARY DATA

Table 2

*Grade 8 rater averages by batch for social studies, writing, and language usage.*

Social Studies

| Scorer ID | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Average |
|---|---|---|---|---|---|---|
| Standard | 0.75 | 0.70 | 0.75 | 0.85 | 0.85 | 0.78 |
| Rater 1 | 0.80 | 0.90 | 1.10 | 0.90 | 0.10 | 0.76 |
| Rater 2 | 0.85 | 0.60 | 1.05 | 0.75 | 0.75 | 0.80 |
| Rater 3 | 0.80 | 0.70 | 0.95 | 0.70 | 0.95 | 0.82 |
| Rater 4 | 0.50 | 0.45 | 0.70 | 0.65 | 0.55 | 0.57 |
| Rater 5 | 0.60 | 0.65 | 0.80 | 0.75 | 0.85 | 0.73 |

Writing

| | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Average |
|---|---|---|---|---|---|---|
| Scorer ID | 1.15 | 0.73 | 0.95 | 0.60 | 1.10 | 0.91 |
| Standard | 1.00 | 0.90 | 1.25 | 0.85 | 0.90 | 0.98 |
| Rater 1 | 1.35 | 0.95 | 1.20 | 1.00 | 1.20 | 1.14 |
| Rater 2 | 0.95 | 0.95 | 0.90 | 0.95 | 1.00 | 0.95 |
| Rater 3 | 0.65 | 0.60 | 0.65 | 0.70 | 1.00 | 0.72 |
| Rater 4 | 0.65 | 0.65 | 0.90 | 0.85 | 1.10 | 0.83 |
| Rater 5 | | | | | | |

Language Usage

| Scorer ID | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Average |
|---|---|---|---|---|---|---|
| Standard | 1.20 | 0.85 | 1.05 | 0.95 | 1.35 | 1.08 |
| Rater 1 | 1.35 | 1.10 | 1.45 | 1.15 | 1.40 | 1.29 |
| Rater 2 | 1.50 | 1.15 | 1.30 | 1.15 | 1.35 | 1.29 |
| Rater 3 | 1.15 | 0.95 | 1.05 | 1.10 | 1.10 | 1.07 |
| Rater 4 | 0.95 | 0.80 | 1.05 | 0.95 | 1.30 | 1.01 |
| Rater 5 | 1.20 | 1.30 | 1.25 | 1.05 | 1.35 | 1.23 |

Note: Batches represent rater averages (n = 20) on scales of 0 - 2.





PRELIMINARY DATA

Table 3

*The standard averages and the aggregate average across all raters for all content areas*

| | | |
|---|---|---|
| **Grade 3** | | |
| Social Studies | | |
| | Standard Average | 0.51 |
| | Aggregate Average | 0.31 |
| Writing | | |
| | Standard Average | 0.73 |
| | Aggregate Average | 0.51 |
| Language Usage | | |
| | Standard Average | 0.68 |
| | Aggregate Average | 0.57 |
| **Grade 8** | | |
| Social Studies | | |
| | Standard Average | 0.78 |
| | Aggregate Average | 0.74 |
| Writing | | |
| | Standard Average | 0.91 |
| | Aggregate Average | 0.92 |
| Language Usage | | |
| | Standard Average | 1.08 |
| | Aggregate Average | 1.18 |





PRELIMINARY DATA